The ICSI-SRI Spring 2006 Meeting Recognition System

Adam Janin   Andreas Stolcke

Xavier Anguera   Kofi Boakye

Özgür Çetin   Joe Frankel

Arindam Mandal   Chuck Wooters   Jing Zheng

International Computer Science Institute, Berkeley, CA, USA

SRI International, Menlo Park, CA, USA

Overview

- Data and Tasks
- Audio preprocessing
  - More robust MDM delay-sum processing
  - Improved IHM cross-talk suppression method
- SRI decoding architecture
- Acoustic modeling
  - MLP feature retraining
  - Pooling of near and distant mic data
  - Adaptation to lecture task
- Language modeling
- Overall results
- Conclusions and future work

Conference Meeting Datasets

- **eval06**: NIST RT-06S conference meetings
- **eval05**: NIST RT-05S conference meetings
  - Unbiased test set
- **dev04a + eval04**: RT-04S devtest and eval meetings plus 2 AMI excerpts
  - Used for development and tuning
  - In spite of lapel mics in CMU and LDC meetings
- Meeting training data (identical to last year!)
  - AMI (35 meetings, 16 hours)
  - CMU (17 meetings, 11 hours) – Lapel personal mics, no distant mics
  - ICSI (73 meetings, 74 hours)
  - NIST (15 meetings, 14 hours)
- Acoustic background training data (same as last year)
  - CTS (Switchboard + Fisher, 2300 hours)
  - BN (Hub-4 + TDT2 + TDT4, 900 hours)

Lecture Datasets

- **eval05:** RT-05S lecture eval set
  - Used for development
  - No independent test set was available
  - Many parameters (e.g., rescoring weights) were copied from conference meeting system without retuning
- **dev06:** Not used (overlapped with eval05)
- Training data:
  - All conference training data
  - Background data as for conference data
  - CHIL training data (close-talking mics only, 38 meetings, ~7 hours)
  - TED lecture recordings (boom mics only, 39 meetings, ~9 hours)

Evaluation Tasks

Conference room meetings:

- **MDM** Multiple distant microphones
- **IHM** Individual headset microphones
- **SDM** Single distant microphone

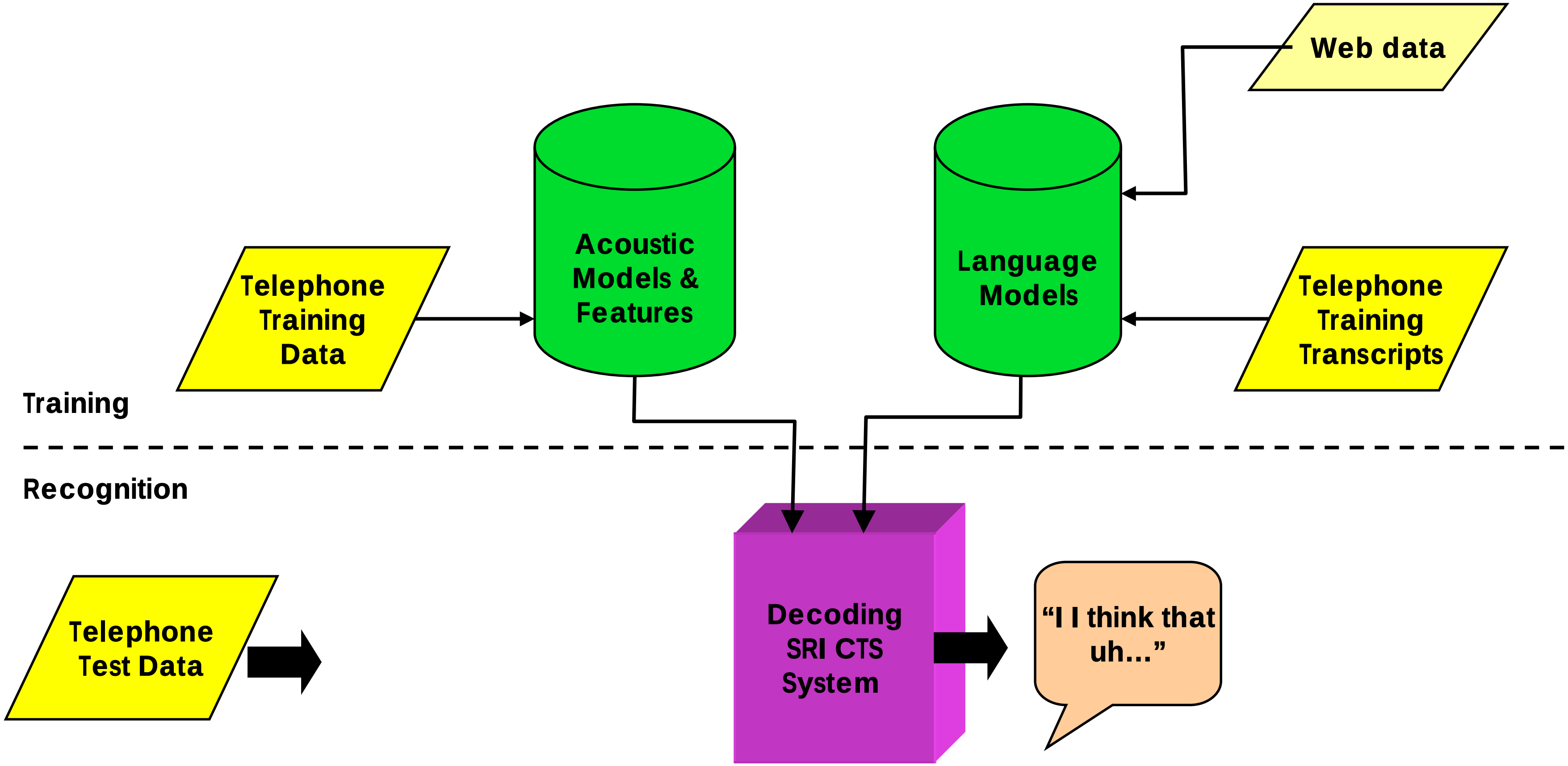
Lecture room meetings (in addition to above):

- **ADM** All distant microphones (i.e., table-top and array)
- **MM3A** Multiple Mark III microphone arrays
  - Precomputed beamformed signal based on UKA source localization estimates
  - Where more than one array was available we only used the first

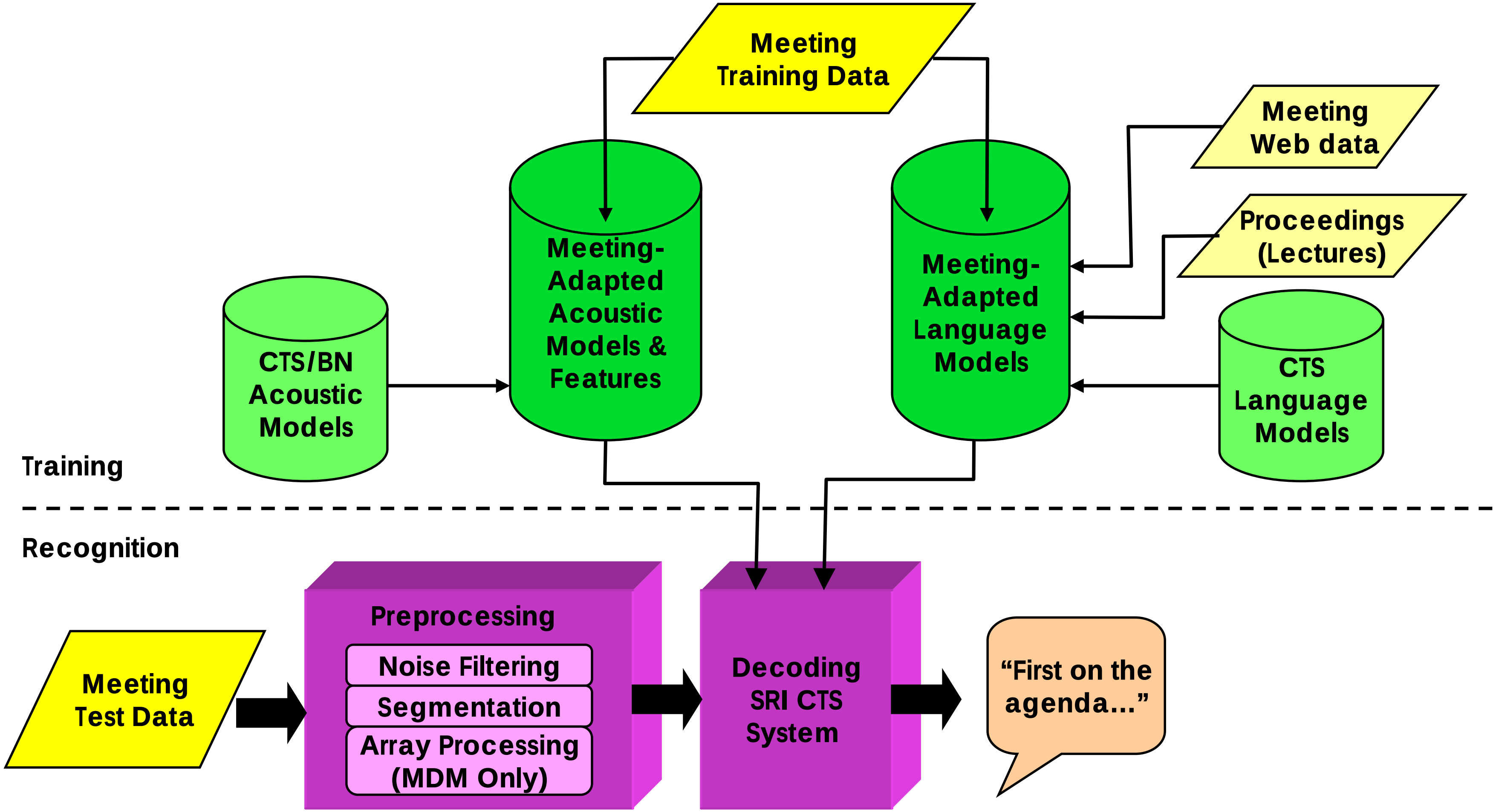
Overlap

- Although the primary evaluation condition was for overlapped speech, all results here are for one speaker only
- No special processing was done to handle overlap

Development Strategy: Base System



Development Strategy: Meeting System



Acoustic Preprocessing

Recognition

- Distant microphones
  - Noise reduction using Wiener filtering on all input channels
  - Delay-sum beamforming of all channels, into single enhanced channel (MDM)
    - *New:* Improved, more robust implementation using a double step viterbi processing algorithm (more details in the diarization talk)
  - Waveform segmentation (speech-nonspeech HMM decoding)
  - Segment clustering (for cepstral normalization, unsupervised adaptation)
- Close-talking (personal) microphones
  - No noise reduction (tried it, no gains)
  - *New:* Waveform segmentation with cross-channel features

Training

- Distant microphones (same as last year)
  - Eliminate overlapping speech (based on personal mic word alignment times)
  - Noise filtering
  - No delay-sum processing
  - Models trained on a selection of distant channel signals

Multiple Distant Microphone Processing

- Results didn’t change significantly for conference meetings
- More robustness on lecture meetings
  - 2005 lecture MDM system was slightly worse than SDM
  - 2006 delay-sum worked well for table-top mics and array mics (MDM and ADM)
  - Most gain comes from array mics
- MDM and ADM now much better than UKA beamformed Mark III signal
  - Caveat: only a single Mark III array was used by our system (but only IBM had more than one Mark III array)
  - The goal of UKA beamformer wasn’t to minimize word-error on this task!

Condition	eval05 lectures	eval06 lectures
SDM	47.7	58.6
MDM	45.8	56.5
ADM	38.6	52.4
UKA/MM3A	n/a	58.3
ICSI/MM3A	n/a	56.9

IHM Crosstalk Suppression

- Last year: energy-based post-processing/filtering of speech detector output
  - Subtract minumum energy (noise floor) from each channel
  - For each channel, subtract average energy of all *other* channels
  - Threshold at zero
  - Intersect foreground segments with HMM speech/nonspeech output

- This year: use cross-channel energy features directly in HMM decoder (along with cepstral features)
  - Min and max log energy difference between target and all non-target channels
  - Trained on first 10 minutes of AMI (35 mtgs), NIST (15 mtgs), and ICSI (73 mtgs)
  - Performed as well as, and more robustly than, cross-correlation based features
  - Eval system used raw energies
  - Post-eval improvement with normalized energies

IHM Crosstalk Suppression

- Using the SDM signal
  - Eval05 included a meeting with an unmiked participant
  - SDM served as “stand-in” mic for participant
  - Including the SDM signal (and energy normalization) improved results by >12% on NIST meetings!
  - The SDM signal was not used for eval06 since there were no unmiked speakers

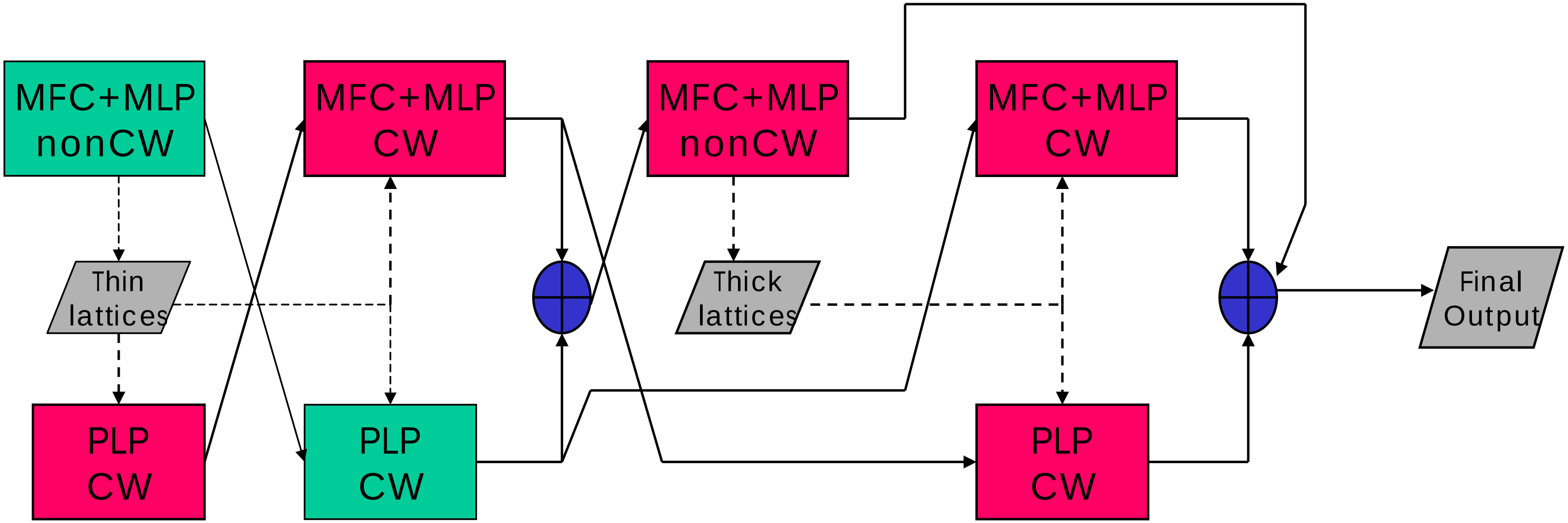
Segmenter Method	**WER**					
	ALL	AMI	CMU	ICSI	NIST	VT
E-diffs (raw)	**25.6**	22.0	23.5	20.9	**37.3**	23.8
E-diffs (raw) +SDM	**24.7**				**33.0**	
E-diffs (norm)+SDM	**22.7**	21.9	23.1	20.6	**25.2**	22.9
Reference	**19.5**	19.2	19.9	16.8	**21.4**	20.6

IHM Segmenter: Conference Results

Segmenter Method	**eval05**				**eval06**
	WER	Sub	Del	Ins	WER
Baseline	**29.3**	11.0	10.3	8.0	
2005 eval system	**25.9**	11.0	11.5	3.4	
2006 eval system (raw energies)	**24.7**	11.1	10.2	3.3	**24.0**
2006 post-eval (normalized energies)	**22.7**	10.9	10.2	1.6	**22.8**
Reference	**19.5**	11.2	6.7	1.6	**20.2**

- 1.2% gain over last year's segmenter on **eval05**
- Energy normalization gave extra 1.2% gain on **eval06**, 2.0% on **eval05** (due to unmiked speaker in NIST meeting)
- Still room for improvement (> 2%) compared to ideal segmentation

SRI System Architecture



MFC+MLP nonCW

MFC+MLP CW

MFC+MLP nonCW

MFC+MLP CW

Thin lattices

Thick lattices

Final Output

PLP CW

PLP CW

PLP CW

Legend

🟩 Decoding/rescoring step

→ Hyps for MLLR or output

- - - -→ Lattice generation/use

▱ Lattice or 1-best output

⊕ Conf. Network combination

Runtime: 12xRT (for CTS, Gaussian shortlists)
25xRT (on meetings, no Gaussian shortlists)

Acoustic Features and Models

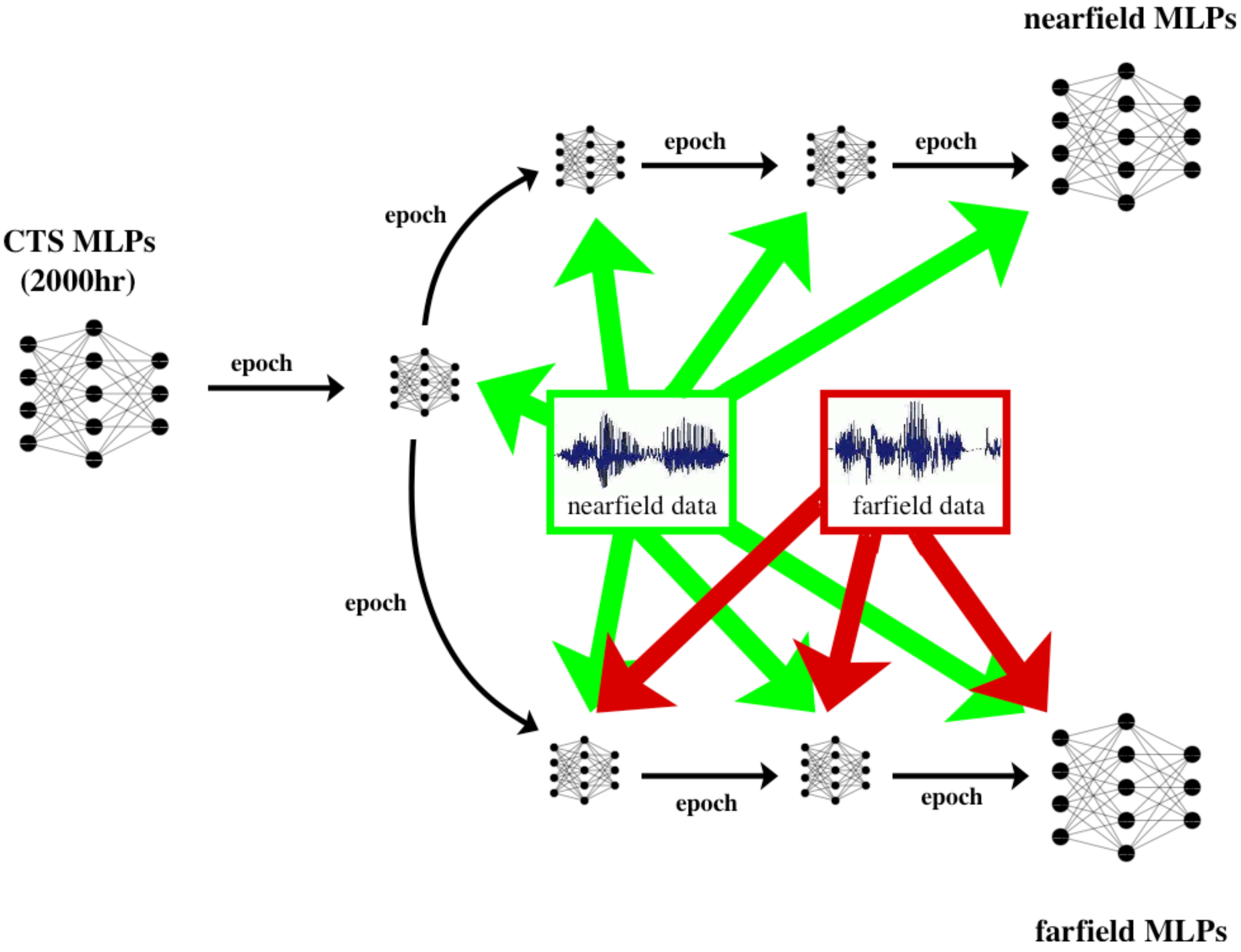
- MFCC within- and crossword triphone models
  - Augmented with 2 x 5 voicing features (5 frames around current frame)
  - Augmented with 25-dim Tandem/HATS phone posterior features estimated by multilayer perceptron (MLP features)
  - Gender dependent
  - Base model trained on 1400h of conversational telephone data
  - MAP adapted to meeting data
- PLP crossword triphone models
  - Gender independent
  - Base models trained on 900h of TDT broadcast data
  - MAP adapted to meeting data
- Normalization and adaptation:
  - CMN + CVN, VTLN
  - HLDA
  - CMLLR (SAT) in training and test (except in first decoding)
  - MLLR with phone-loop in first MFCC and PLP decoding
  - MLLR cross-adaptation in subsequent steps

Acoustic Modeling Changes

Compared to 2005 system

- PLP models now based on BN training for all systems
  - Last year IHM PLP models were CTS-based
  - This change actually gave small degradation on conference meetings
  - … but significant gain on lectures, and more uniformity across all systems
- Distant mic models trained on combination of distant and near mic data
  - Last year: only distant mic data
  - Improves “robustness” to highly effective delay-sum signal enhancement
- Decision-tree based MLLR regression classes
  - Last year: hand-defined regression classes
- MLP features were retrained
  - Adapted to conference meeting data, including AMI
  - Separate MLPs trained for near and distant mic data
  - Last year: near-mic feature adaptation only, used unchanged for distant mics

Posterior Based Features



Details on Posterior Features

- CTS MLPs trained on 8kHz data
- Tandem
  - 3-layer 9-frame PLP input
- Hats
  - 15 critical-band MLPs with 51 frame input
  - Merger net using hidden activations
  - Only merger net was adapted
- 4 epochs of adaptation
- Learning rate equal to the final learning rate of the CTS nets.
- Farfield adaptation only of non-overlap regions (alignments generated from near-field signal)
- Only one farfield channel (chosen at random) was used.

Acoustic Modeling Improvements

- Aggregate improvements
- Excluding changes in segmentation and delay-sum

Models	eval05 Conference		eval05 Lecture	
	SDM	IHM	SDM	IHM
2005	40.9	24.7	51.9	30.8
2006	39.3	24.1	47.7	28.6

Conference Meeting LMs

- Linearly interpolated mixture N-gram LMs
  - Multiword bigram for lattice generation
  - Multiword trigram for lattice decoding
  - Word-based 4-gram for rescoring

  All LMs entropy-pruned
- Conference meeting LM components

  (1) Switchboard CTS transcripts (6.5M words)

  (2) Fisher CTS (23M)

  (3) Hub4 and TDT4 BN transcripts (140M)

  (4) AMI, CMU, ICSI, and NIST meeting transcripts (1M)

  (5) Web data selected to match Fisher (530M) transcripts and meeting (382M) transcripts (newly collected using a new selection criterion)
- Perplexity optimized on held-out data (AMI, CMU, ICSI, NIST)
- Vocabulary: 54K words
  - All words in Switchboard, RT-04S meeting transcripts
  - All non-singletons in Fisher, AMI devtest
  - OOV rates: 0.40% on eval04; 0.19% on AMI devtest

Lecture Meeting LMs

- Similar to conference meeting LM, but:
  - Added CHIL transcripts (70K words)
  - Speech conference proceedings (32M)
  - Removed Fisher web data
  - Collected new web data based on CHIL transcripts (512M)
- Vocabulary: added 3781 words from conference proc.
  - OOV rate on CHIL devtest: 0.18%
  - Most common OOV word in CHIL: ixy
- Perplexity optimized on a portion of the CHIL training data
- As compared to the 2005 lecture LMs, a sizeable reduction in perplexity (5%) but only a small improvement in WER (0.1)

Conference Meetings: Overall Results

- Relative WER reduction on eval05 data:
  - 3.9% for MDM and SDM
  - 4.0% for IHM
- Additional gain > 1% for IHM after evaluation (improved segmenter)
- eval06 difficulty similar to eval05
  - Possible exception: MDM (more AMI data?)

System	MDM	SDM	IHM
	eval05		
RT-05S	**30.2**	**40.9**	**25.9**
RT-06S (post-eval)	**29.0**	**39.3**	**24.1** **(23.0)**
	eval06		
RT-06S (post-eval)	**34.2**	**41.2**	**24.1** **(22.8)**

Other Lecture System Differences

- Based on conference meeting system
- Acoustic models derived from conference meeting models with an additional adaptation step using (on eval05):
  - CHIL data (1.2% improvement on IHM, not used on ADM)
  - TED data (0.5% improvement on IHM, 0.6% on ADM)
- No distant mic data in CHIL training set, therefore adapted eval06 models to dev06 data
  - But not for eval05 experiments due to speaker overlap!
- Model score weights not optimized
  - Used parameters optimized for conference meetings
- Energy normalization in IHM segmenter does not help (actually hurts), maybe due to gain issues
- Speaker clustering for distant mic recognition does not help; use a single speaker cluster instead

Lecture Recognition: Overall Results

- Large improvements on 2005 eval data
- 2006 eval data was much harder
- Possible factors:
  - Mostly nonnative speakers
  - More recording sites, therefore more variation in recording conditions
  - More channels in IHM condition (high insertion rates from crosstalk)

System	MDM	ADM	MM3A	SDM	IHM
	eval05				
RT-05S	52.0	44.8	-	51.9	28.0
RT-06S	45.8	38.6	-	47.7	23.8
	eval06				
RT-06S	56.5	52.4	58.3 / 56.9	58.6	49.6

Conclusions

- Little change in conference meeting system
  - Modest gains
- Significantly improved IHM segmenter
  - Integration of cepstral and cross-channel energy features
- Big improvements in lecture recognition system
  - Use of conference-trained distant mic MLP features
  - Combination of CTS and BN models
  - More robust delay-sum
  - Use of CHIL and TED data to adapt base models
  - Small LM improvement
  - Our first serious effort for lecture meeting recognition, but all done within a couple weeks before and during the evaluation!

Future Work

Last year's List with Things That Got Done

- Fix the things we didn't have time for
  - MMI-MAP for distant mic models
  - Adapt MLPs for distant mic features √
- Solve the IHM segmentation problem! √ (sort of)
- Do a better job on lecture recognition
  - Use TED acoustic data √
  - Get MDM to work better than SDM √
  - Estimate model weights (LM weight, insertion penalty, …) properly
- Explore feature mapping techniques (e.g. MLLR) to reduce mismatch of background training data
- Adapt model to non-native speakers of Amer. English
  - Germans, Brits, Scots, …
  - cf. Arlo Faria's poster, MLMI-05
- More generally, more adaptation to meeting type & content

New Future Work

- Overlap detection and processing

Thank You!